IV Estimation and its Limitations:

Weak Instruments and Weakly Endogeneous Regressors

Laura Mayoral, IAE, Barcelona GSE and University of Gothenburg

U. of Gothenburg, May 2015

Roadmap

- Testing for deviations from the standard framework
- Detection of weak instruments
- Inference robust to weak instruments
- Software: ivreg2 (STATA)
- Examples

Testing for deviations from the standard framework

- Testing for the failure of the exclusion restriction: not possible to test, unless the model is overidentified (i.e., there are more instruments that endogenous variables), but strong assumptions are needed (basically you need to assume what you want to test!)

- Testing for the failure of the relevance condition/weak instruments: different tests are available, most of them based on the F statistic of the null hypothesis that $\Pi = 0$ in the first stage.

Testing the exclusion restriction: J tests of overidentifying restrictions

- In an exactly identified model we cannot test the hypothesis that the instrument is valid, i.e. that the exclusion restriction is a valid one$\longrightarrow$ the assumption that the instrument is valid will essentially have to be taken on faith.

- See Murray (2005, 2006) for tips that will help you to motivate this "faith"

- If the model is overidentified, it's possible to test for the validity of the overidentifying restrictions.

- Strong assumption: you need to assume that a set of instruments (enough to identify the parameters) are valid and you can test whether the remaining ones are also one.

- If you reject the null that the remaining instruments are exogenous: there is something wrong in your model–you need to reconsider the validity of the instruments

- But not rejecting the null doesn't mean that your instruments are valid!!

- Sargan test: you don't need to specify which instruments are the valid ones and which are the dubious ones.

- But still, you need as least K instruments that are valid! (K=number of endogenous variables). Otherwise the test is not consistent.

Detection of weak instruments

- Two types of test

- Tests of underidentification: test whether instruments are irrelevant (Π=0)

- Tests of weak identification: test whether instrument are weak.

- Always interpret tests of underidentification with caution:

- if you reject underidentification, it can still be the case that your model is only weakly identified since instruments are weak.

Detection of weak instruments, II

- Goal: determine when instruments are irrelevant (test of under identification) or weak (tests of weak identification).

- For the 1 endogenous variable case, consider the reduced-form regression

$$X = Z\Pi + W\delta + v$$

where W are exogenous regressors.

Using F to test for weak instruments

- The concentration parameter $\mu^2 = \Pi' Z' Z \Pi / \sigma_v^2$ is closely related to the F statistic associated to testing $H_0 : \Pi = 0$.

- More specifically (under conditional homocedasticity) $F \simeq 1 + \mu^2 / K$, where K is the number of instruments.

- weak instruments $\longrightarrow$ a low value of μ^2 $\longrightarrow$ a low value of F.

- But: this relation relies heavily on the assumption of conditional homokedasticity of the error term.

Staiger and Stock's rule of thumb

- Staiger and Stock (1997) showed that the weak instrument problem can arise even if the hypothesis of $\Pi = 0$ is rejected at standard t-tests at conventional significance levels.

- Staiger and Stock's Rule of thumb (1 endogenous variable):

Reject that your instruments are weak if F$\geq$10,

where F is the F-statistic testing Π=0 in the regression of y on Z and W (where W are the exogenous regressors included in the equation of interest).

Stock and Yogo (2005)'s bias and size methods

- Stock and Yogo (SY, 2005) formalise Staiger and Stock's procedure.

- They show that their rule of thumb is not always enough! (see below)

- SY's tests can be used with multiple endogenous regressors and multiple instruments.

How weak is weak?

- We need a cutoff value for μ^2 such that for values smaller than μ^2, instruments are deemed to be weak.

- Since (under conditional homokedasticity) the F statistic of the first stage regression and μ^2 are related, the cutoff value is usually compared with it.

- There are different ways in which we can define this cutoff value, depending what we want to control (bias of β, size of t-test), etc.

- The value of the cutoff depends on the method employed.

- This means that the same instrument can be weak for one estimation method but not for other!

- 2SLS is one of the least robust

How weak is weak, II

- Stock and Yogo propose two methods to select the cutoffs:

- The bias method: find a cutoff value for μ^2 such that μ^2 the relative bias of the 2SLS with respect to OLS doesn't exceed certain quantity.

- The size method: find a cutoff value for μ^2 such that the maximal size of a Wald test of all the elements of β doesn't exceed a certain amount.

Stock and Yogo (2005)'s bias method

- Let $\mu^2_{10\%bias}$ be the value of μ^2 such that if $\mu^2 \geq \mu^2_{10\%bias}$, the maximum bias of the 2SLS estimator will be no more than 10% of the bias of OLS.

- Decision rule is (5% significance level)

Reject that your instruments are weak if $F > J_{10}(k)$,

where $J_{10}(k)$ is chosen such that $P(F > J_{10}(k); \mu^2 = \mu^2_{10\%bias}) = 0.05$

Stock and Yogo (2005)'s bias method, II

- Test statistic:

- 1 endogenous regressor: computed using F from the first stage regression.

- more than 1 endogenous regressor: Cragg- Donald (1993) statistic (multivariate version of the F statistic).

- Stock and Yogo (2005) provide critical values that depend on

- the number of endogenous regressors,
- the number of instruments,
- the maximum bias.
- The estimation procedure (2SLS, LIML, ...)

- ivreg2 (STATA): gives you the critical values.

Stock and Yogo (2005)'s size method

- Similar logic but instead of controlling bias, control the size of a Wald test of $\beta=\beta_0$

- Use F statistic first stage if 1 endogenous regressor or Cragg-Donald if more than 1

- Decision rule:

- Reject weak instruments if statistic is larger than the critical value

- Have a look at the critical values (and compare to Staiger and Stock's rule of thumb!)

Table 5.1. *Critical values for the weak instrument test based on TSLS bias* (Significance level is 5%)

	$n = 1, b =$				$n = 2, b =$				$n = 3, b =$			
K_2	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
3	13.91	9.08	6.46	5.39								
4	16.85	10.27	6.71	5.34	11.04	7.56	5.57	4.73				
5	18.37	10.83	6.77	5.25	13.97	8.78	5.91	4.79	9.53	6.61	4.99	4.30
6	19.28	11.12	6.76	5.15	15.72	9.48	6.08	4.78	12.20	7.77	5.35	4.40
7	19.86	11.29	6.73	5.07	16.88	9.92	6.16	4.76	13.95	8.50	5.56	4.44
8	20.25	11.39	6.69	4.99	17.70	10.22	6.20	4.73	15.18	9.01	5.69	4.46
9	20.53	11.46	6.65	4.92	18.30	10.43	6.22	4.69	16.10	9.37	5.78	4.46
10	20.74	11.49	6.61	4.86	18.76	10.58	6.23	4.66	16.80	9.64	5.83	4.45
11	20.90	11.51	6.56	4.80	19.12	10.69	6.23	4.62	17.35	9.85	5.87	4.44
12	21.01	11.52	6.53	4.75	19.40	10.78	6.22	4.59	17.80	10.01	5.90	4.42
13	21.10	11.52	6.49	4.71	19.64	10.84	6.21	4.56	18.17	10.14	5.92	4.41
14	21.18	11.52	6.45	4.67	19.83	10.89	6.20	4.53	18.47	10.25	5.93	4.39
15	21.23	11.51	6.42	4.63	19.98	10.93	6.19	4.50	18.73	10.33	5.94	4.37
16	21.28	11.50	6.39	4.59	20.12	10.96	6.17	4.48	18.94	10.41	5.94	4.36
17	21.31	11.49	6.36	4.56	20.23	10.99	6.16	4.45	19.13	10.47	5.94	4.34
18	21.34	11.48	6.33	4.53	20.33	11.00	6.14	4.43	19.29	10.52	5.94	4.32
19	21.36	11.46	6.31	4.51	20.41	11.02	6.13	4.41	19.44	10.56	5.94	4.31
20	21.38	11.45	6.28	4.48	20.48	11.03	6.11	4.39	19.56	10.60	5.93	4.29
21	21.39	11.44	6.26	4.46	20.54	11.04	6.10	4.37	19.67	10.63	5.93	4.28
22	21.40	11.42	6.24	4.43	20.60	11.05	6.08	4.35	19.77	10.65	5.92	4.27
23	21.41	11.41	6.22	4.41	20.65	11.05	6.07	4.33	19.86	10.68	5.92	4.25
24	21.41	11.40	6.20	4.39	20.69	11.05	6.06	4.32	19.94	10.70	5.91	4.24
25	21.42	11.38	6.18	4.37	20.73	11.06	6.05	4.30	20.01	10.71	5.90	4.23
26	21.42	11.37	6.16	4.35	20.76	11.06	6.03	4.29	20.07	10.73	5.90	4.21
27	21.42	11.36	6.14	4.34	20.79	11.06	6.02	4.27	20.13	10.74	5.89	4.20
28	21.42	11.34	6.13	4.32	20.82	11.05	6.01	4.26	20.18	10.75	5.88	4.19
29	21.42	11.33	6.11	4.31	20.84	11.05	6.00	4.24	20.23	10.76	5.88	4.18
30	21.42	11.32	6.09	4.29	20.86	11.05	5.99	4.23	20.27	10.77	5.87	4.17

Notes. The test rejects if $g_{\min}$ exceeds the critical value. The critical value is a function of the number of included endogenous regressors (n), the number of instrumental variables (K_2), and the desired maximal bias of the IV estimator relative to OLS (b).

Table 5.2. *Critical values for the weak instrument test based on TSLS size* (Significance level is 5%)

	$n = 1, r =$				$n = 2, r =$			
K_2	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	19.93	11.59	8.75	7.25	7.03	4.58	3.95	3.63
3	22.30	12.83	9.54	7.80	13.43	8.18	6.40	5.45
4	24.58	13.96	10.26	8.31	16.87	9.93	7.54	6.28
5	26.87	15.09	10.98	8.84	19.45	11.22	8.38	6.89
6	29.18	16.23	11.72	9.38	21.68	12.33	9.10	7.42
7	31.50	17.38	12.48	9.93	23.72	13.34	9.77	7.91
8	33.84	18.54	13.24	10.50	25.64	14.31	10.41	8.39
9	36.19	19.71	14.01	11.07	27.51	15.24	11.03	8.85
10	38.54	20.88	14.78	11.65	29.32	16.16	11.65	9.31
11	40.90	22.06	15.56	12.23	31.11	17.06	12.25	9.77
12	43.27	23.24	16.35	12.82	32.88	17.95	12.86	10.22
13	45.64	24.42	17.14	13.41	34.62	18.84	13.45	10.68
14	48.01	25.61	17.93	14.00	36.36	19.72	14.05	11.13
15	50.39	26.80	18.72	14.60	38.08	20.60	14.65	11.58
16	52.77	27.99	19.51	15.19	39.80	21.48	15.24	12.03
17	55.15	29.19	20.31	15.79	41.51	22.35	15.83	12.49
18	57.53	30.38	21.10	16.39	43.22	23.22	16.42	12.94
19	59.92	31.58	21.90	16.99	44.92	24.09	17.02	13.39
20	62.30	32.77	22.70	17.60	46.62	24.96	17.61	13.84
21	64.69	33.97	23.50	18.20	48.31	25.82	18.20	14.29
22	67.07	35.17	24.30	18.80	50.01	26.69	18.79	14.74
23	69.46	36.37	25.10	19.41	51.70	27.56	19.38	15.19
24	71.85	37.57	25.90	20.01	53.39	28.42	19.97	15.64
25	74.24	38.77	26.71	20.61	55.07	29.29	20.56	16.10
26	76.62	39.97	27.51	21.22	56.76	30.15	21.15	16.55
27	79.01	41.17	28.31	21.83	58.45	31.02	21.74	17.00
28	81.40	42.37	29.12	22.43	60.13	31.88	22.33	17.45

Detecting weak instruments, final comments

- As mentioned before, the logic of using the first-stage using the F statistics relies heavily on the assumption of conditional homokedasticity.

- Solution: ongoing research

- Kleibergen-Paap rk Wald statistic: ivreg2 reports this test as a test for weak instruments when robust options are called for.

- However, this test is not formally justified in the context of weak instruments.

- It is justified in the case of under identification and if errors are i.i.d., it becomes the Cragg-Donald test (but not under weak instruments!).

- Olea Montiel and Pflueger (2013) and (2014): tests valid under heteroskedasticity, autocorrelation, and clus- tering robust weak instrument tests, for 2SLS and LIML.

- But only applicable if there is only 1 endogenous regressor.

- Stata: weakivtest (you need to install the package first).

- See also Andrews 2014

- Angrist and Pischke (2009) introduced first-stage F statistics for tests of under- and weak identification when there is more than one endogenous regressor.

- In contrast to the Cragg-Donald and Kleibergen-Paap statistics, which test the identification of the equation as a whole, the AP first-stage F statistics are tests of whether one of the endogenous regressors is under- or weakly identified.

Inference methods robust to weak instruments

- Problem: is it still possible to test hypothesis about β if instruments are weak?

- YES!

- But, why do you want to do that?

- If you know your instruments are weak, you should look for better instruments!

- But there are cases where the literature on weak IV detection is still limited (non i.i.d. errors for instance) so maybe you don't know.

- Two approaches to improving inference:

- Fully robust methods: Inference that is valid for any value of the concentration parameter (including zero!), at least if the sample size is large, under weak instrument asymptotics.

  - For tests: asymptotically correct size (and good power!)

  - For confidence intervals: asymptotically correct coverage rates

- Partially robust methods: Methods are less sensitive to weak instruments than 2SLS (e.g. bias is "small" for a "large" range of values of μ^2)

Fully Robust Testing

- Results are much more developed for the case of 1 endogenous regressor

- If more than 1 endogenous regressor: still an open econometric problem.

- The model (1 endogenous regressor).

$$y = X\beta + \epsilon$$

where X is $(N \times 1)$ and endogenous, Z is a matrix of instruments, maybe weak. We want to test $H_0 : \beta = \beta_0$.

- Two approaches

- Approach 1: use a statistic whose distribution does not depend on μ^2

  - Two statistics here: The Anderson-Rubin (1949) test and LM statistic (Moreira 2002 and Kleiberger 2002).

- Approach 2: Use statistics whose distribution depends on μ^2, but compute the critical values as a function of another statistic that is sufficient for μ^2 under the null hypothesis.

  - Conditional likelihood ratio test (Moreira 2003).

- Both approaches have advantages and disadvantages. We discuss both.

Tests robust to Weak instruments: Approach 1

- The Anderson-Rubin test

Set up: $y = X\beta + \epsilon$, X endogenous, Z is a matrix $(N \times k)$ of instruments. We want to test $H_0 : \beta = \beta_0$.

- AR's test: F test in the regression of $(y - X\beta_0)$ on Z.

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_z / (y - X\beta_0) / k}{(y - X\beta_0)' M_z (y - X\beta_0)} / (N - k)$$

where $P_z = Z'(Z'Z)^{-}1Z$ and $M_z = I - P_z$

- under H_0, $y - X\beta_0 = \epsilon$, so (if ϵ is iid)

$$AR(\beta_0) = \frac{\epsilon' P_z \epsilon / k}{\epsilon' M_z \epsilon / (N - k)} \xrightarrow{d} \chi^2_k / k$$

Advantages and disadvantages of AR

- Advantages

- Easy to use (entirely regression based)

- Uses standard F critical values (asymptotic distribution doesn't depend on μ^2

- Works for m $>$ 1 (more than 1 endogenous regressor).

- Disadvantages

- Difficult to interpret: rejection arises for two reasons: H_0 is false or Z is endogenous!

- Power loss relative to other tests (we shall see)

- Is not efficient if instruments are strong –under strong instruments, not as powerful as TSLS Wald test—.

Approach 1, cont. Kleibergen's (2002) LM test

- Kleibergen developed an LM test that has a null distribution that is also χ^2_1 (doesn't depend on μ^2)

- Advantages
  - Fairly easy to implement
  - Is efficient if instruments are strong

- Disadvantages
  - Has very strange power properties (power function isn't monotonic)
  - Its power is dominated by the conditional likelihood ratio test

Tests robust to Weak instruments. Approach 2: Conditional maximum likelihood tests

- Recall your probability and statistics courses:
- Let S be a statistic with a distribution that depends on θ
- Let T be a sufficient statistic for θ
- Then the distribution of $S|T$ does not depend on θ

Conditional maximum likelihood tests

- Moreira (2003):

- LR will be a statistic testing $\beta = \beta_0$ (LR is "S" in notation above)

- Q_T will be sufficient for μ^2 under the null (Q_T is "T")

- Thus the distribution of $LR|Q_T$ does not depend on μ^2 under the null

- Thus valid inference can be conducted using the quantiles of $LR|Q_T$; that is, using critical values that are a function of Q_T

- Implementation: condivreg (STATA)

- Advantages
  - More powerful than AR or LM

- Disadvantages
  - More complicated to explain and write down
  - Only developed (so far) for a single included endogenous regressor
  - As written, the software requires homoskedastic errors; extensions to heteroskedasticity and serial correlation have been developed but are not in common statistical software

Constructing confidence intervals

- It is possible to construct confidence intervals for β by inverting the robust tests described above.

- How?

- Test all the hypotheses of the form $H_0 : \beta = \beta_0$ for different values of β_0
- Examine the set of values for which H_0 could not be rejected.

- Inverting the AR test: see Dufour and Taamouti (2005).

- Inverting the Conditional test: see Mikusheva (2005)–only available for 1 endogenous regressor.

- the CI based on conditional test is more efficient than that based on AR.

- Extensions: More than 1 endogenous regressor.

- This literature is still very incomplete

- Kleiberger (2007) provides an extension of the AR test when the interest in testing a joint hypotheses on β.

Partially Robust Estimators

- Estimation under weak instruments is much harder than testing or confidence intervals

- Estimation must be divorced from confidence intervals (ie., use robust methods!).

k-class estimators

$$\hat{\beta} = [X'(I - k^* M_Z X)^{-}1[X'(I - k^* M_Z)y]$$

- 2SLS: $k^* = 1$

- LIML: $k^* = k_{liml}$ (smallest root of some matrix A)

- 2SLS: $k^* = k\text{liml} - H$, (where H is a function of exogenous regressors)

- Under strong instruments, LIML, 2SLS and Fuller will be similar

- Under weak instruments, 2SLS has greater bias and larger MSE

- LIML has the advantage of minimizing the AR statistic–thus, will be always contained in the AR (and CLR) confidence set.

- These properties make LIML be a reasonable good choice as an alternative to 2SLS

Wrapping up

1. Always test formally for weak instruments (but be aware from the limitations of the existing theory)

2. Divorce estimation and testing (i.e., it's not enough to test the significance of β using the standard errors of the estimators.)

3. Conduct hypothesis testing using the conditional likelihood ratio (CLR) test or its robust variants.

4. Build confidence intervals based on CLR.